AD-A230 165

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 26, 1990	Final Report, 1 Jun 88 to 30 Sep 90

4. TITLE AND SUBTITLE
PROBABILISTIC ANALYSIS OF NEURAL NETWORKS

5. FUNDING NUMBERS
AFOSR-88-0189
61102F 2304/A7

6. AUTHOR(S)
William G. Faris
Charles M. Newman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Program in Applied Matheamtics
University of Arizona
Tucson, AR 85721

AFOSR-TR. 90 1154

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
Bldg 410
Bolling AFB DC 20332-6448

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFOSR-88-0189

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release;
distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The research was a probabilistics study of neural network models. It was not oriented toward the workings of a particular device, but was intended to provide an understanding of the basic mechanisms of learning and recognition in neural networks. The main areas of progress were analysis of neural network models, study of network connectivity, and investigation of computer network theory.

14. SUBJECT TERMS	15. NUMBER OF PAGES
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

Probabilistic Analysis of Neural Networks
AFOSR Grant No. 88–0189
Final Technical Report

William G. Faris
Charles M. Newman
Program in Applied Mathematics
University of Arizona
Tucson, Arizona 85721

November 26, 1990

Abstract

The research was a probabilistic study of neural network models. It was not oriented toward the workings of a particular device, but was intended to provide an understanding of the basic mechanisms of learning and recognition in neural networks. The main areas of progress were analysis of neural network models, study of network connectivity, and investigation of computer network theory.

Contents

1 Introduction

The research dealt with probabilistic analysis of neural networks. The principal investigators were William G. Faris and Charles M. Newman. Robert S. Maier also participated in the project.

The analysis was intended to provide an understanding of the basic mechanisms of the elementary components of neural network recognition and memory devices. The research was less concerned with building practical devices than with a mathematical analysis of the basic phenomena. However the analyses will eventually be useful in understanding the components of much more complicated realistic devices.

There were three main components of research:

- Neural Network Models
- Network Connectivity
- Computer Networks

These components are explained in the following sections.

2 Neural Network Models

2.1 Reliability of neural networks

The concept of a set of *associated* random variables provides a nonlinear generalization of the notion of a set of positively correlated random variables. This concept is useful in analyzing the reliability of neural network associative memory devices. In particular, it allows estimates of the probability of successful retrieval of individual bits in a memorized pattern to be extended to a lower bound on the probability of error-free retrieval of the pattern as a whole [1].

There are a number of standard neural network models of memory in which associations are stored in connections between nodes. In order to analyze the problem of when such a memory device can function reliably and when it becomes overloaded, it is useful to treat the memorized items as random. However when this is done, the presence or absence of the various

possible connections are not independent random events. A rigorous analysis must use some concept that goes beyond independence.

The notion of *associated* random variables is one such tool. A set of associated random variables has the property that every pair of variables is positively correlated. However the notion is stronger, since the same property also holds for certain nonlinear functions of all the variables. Many of the properties of independent random variables hold for associated random variables, notably the central limit theorem. (Another terminology has arisen in statistical mechanics; associated random variables are said to obey the FKG inequalities.)

The model is an m-by-n matrix of connections in which pairs of patterns are stored, and from which an output pattern may be retrieved upon presentation of the corresponding input pattern. Each input pattern is a random pattern of n 0's and 1's, the probability of a 1 being p. Similarly, each output pattern is a random pattern of m 0's and 1's, the probability of a 1 being q. The entries in the matrix are also 0's and 1's, and are initially all 0's. Whenever a pattern pair is stored, matrix elements are set to 1 (*i.e.*, are "activated") if there are 1's in the corresponding input and output lines (*i.e.*, the lines are "active"). In all there are z input-output pairs to be stored.

The retrieval process is governed by a threshold parameter λ. If a previously stored input pattern is presented to the matrix, the retrieved output pattern will by definition have 1's on the output lines that are connected by the matrix to a number of active input lines that exceeds the threshold. Under restrictions on the parameters p, q, and z that ensure that the mean density of connections is less than unity, the threshold λ may be chosen so that one gets a good lower bound for the probability that the retrieved pattern equals the originally stored pattern. That is to say, the model parameters can be chosen so that there is minimal interference between the stored pairs of patterns.

This model may have biological interest as a model for associative memory, for instance in the hippocampus. Similar robust models may be important for neural computing.

2.2 Evaluation of neural networks

The question considered is that of reliable evaluation of a neural network. We investigate when the behavior of a neural network in a limited number

of random trials gives average results that are representative of the average results for infinitely many trials [2].

A feed-forward neural net architecture defines a class of functions F_θ (say from $\mathbf{R}^d$ to $\mathbf{R}$) for θ in some parameter set (of weights). We consider n random inputs $\mathbf{X}_i$, for $i = 1, \ldots, n$, taken independently from some probability distribution. The weak law of large numbers gives for each θ a bound on the probability of the event that the sample average differs from the expectation by $\epsilon > 0$.

The research gives a bound on the probability that there exists a θ such that this happens. This is a uniform large deviation result that is considerably more subtle.

One situation where this is important is when the parameter values Θ are random, perhaps depending on the $\mathbf{X}_1, \ldots, \mathbf{X}_n$. In that case the long range expectation should be computed by taking an independent copy $\tilde{\mathbf{X}}$ of $\mathbf{X}$ and computing the conditional expectation of $F_\Theta(\tilde{\mathbf{X}})$ given $\mathbf{X}_1, \ldots, \mathbf{X}_n$. The preceding result then gives the same bound on the probability that the sample average with the random parameter differs from the expectation by ϵ. Thus if a certain laboriously acquired sample is used for the learning trials that define the network, then the same sample may be used for later trials that evaluate its performance.

Such a measurement of performance is relevant to classification problems in which there is more than one probability distribution. The network is intended to distinguish between these distributions. It performs well in the long run if the expectations for the distributions are far apart. The result says that taking moderate size samples from each of the distributions gives a good idea of the corresponding expectations.

The application is to a feed-forward neural network specifying such a function. In the usual terminology of neural networks there is an input layer, a hidden layer, and an output layer. Associated with the layers are d input nodes, N hidden nodes, and one output node. The same non-linear *threshold function* ϕ from $\mathbf{R}$ to $\mathbf{R}$ is associated with each hidden node. It is customary to take it to be continuous and increasing from 0 to 1.The threshold function is fixed in advance. One conventional choice is $\phi(u) = 1/(1 + e^{-u})$. Often there is also a threshold function associated with output nodes. Since this merely amounts to a change of variable, we omit this complication.

The network is specified during the learning process by assigning connection weights. For each hidden layer node i there is a scalar a_i representing a

bias weight and a vector $\mathbf{b}_i$ in $\mathbf{R}^d$ corresponding to the weights coming from the input nodes. Also for each hidden layer node i there is a scalar c_i representing the weight going to the output node. We specify the network in a compact way by giving the vector $\mathbf{a}$ of biases, the matrix $\mathbf{B}$ of input weights, and the vector $\mathbf{c}$ of output weights. (The output weights are bounded by a fixed constant.) These quantities constitute the parameter space.

The reason for this choice of architecture is the observation that it is sufficient for synthesizing arbitrary functions. That is, a network of this type with sufficiently many intermediate nodes is able to approximate most functions rather well.

The uniform large deviation result applies to this sort of neural network. The probability estimate involves a certain polynomial p_d of degree d. (For example $p_1(m) = 2m$ and $p_2(m) = m^2 - m + 2$.) The statement is about a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of independent and identically distributed random vectors in $\mathbf{R}^d$. It says that for every $\epsilon > 0$ there exists $\epsilon' > 0$ such that for large n the probability that there exists $\mathbf{a}, \mathbf{B}, \mathbf{c}$ such that the sample average differs from the expectation by more than ϵ is bounded by $4p_d(2n)\exp(-(\epsilon')^2 n/8)$. The important point is that as $n \to \infty$ the exponential goes to zero faster than the polynomial goes to infinity. The probability of a large deviation is thus very small.

2.3 Neural network loss functions

The work on reliable evaluation was extended to encompass a global measure of loss [3].

A neural network is a system that is supposed to perform its function by learning from experience rather than being programmed by an algorithm. The class of network architectures that we consider is representative. In considering this class we fix a smooth non-linear threshold function ϕ from $\mathbf{R}$ to $\mathbf{R}$. We next choose integers d and N, corresponding to the number of input nodes and the number of hidden nodes. There is a single output node. These choices define the architecture.

The network is then specified by parameters called connection weights. These are a vector $\mathbf{a}$ in $\mathbf{R}^d$ of biases, an N by d matrix $\mathbf{B}$ of input weights, and a vector $\mathbf{c}$ in $\mathbf{R}^N$ of output weights. We write $\theta = (\mathbf{a}, \mathbf{B}, \mathbf{c})$ for an element of the parameter space.

Let ϕ^N be the function from $\mathbf{R}^N$ to $\mathbf{R}^N$ defined by pointwise evaluation.

The network is then the function from $\mathbf{R}^d$ to $\mathbf{R}$ given by $F_\theta(\mathbf{x}) = \mathbf{c} \cdot \phi^N(\mathbf{B}\mathbf{x} + \mathbf{a})$. The picture is that the transformation from the input nodes to the the hidden nodes is linear, but then there is a thresholding operation at each hidden node. The transformation to the output node is again linear.

Assume that the input to the network is a random vector $\mathbf{X}$ in $\mathbf{R}^d$. A theoretical measure of the performance of a particular network in approximating a function f is the expectation

$$L(\theta) = \mathbf{E}[(f(\mathbf{X}) - F_\theta(\mathbf{X}))^2]. \tag{1}$$

When this is small the performance is good. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent random vectors. The corresponding empirical measure of performance is the random variable

$$\hat{L}_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} (f(\mathbf{X}_k) - F_\theta(\mathbf{X}_k))^2. \tag{2}$$

The weak law of large numbers says that for each θ and for each $\epsilon > 0$ the probability $\mathbf{P}[|\hat{L}_n(\theta) - L(\theta)| > \epsilon] \to 0$ as $n \to \infty$. In many situations there is also a large deviation bound that says that the probability approaches zero exponentially as $n \to \infty$. Our purpose is to present a large deviation bound that is uniform in θ. That is, we show that

$$\mathbf{P}[\exists \theta \, |\hat{L}_n(\theta) - L(\theta)| > \epsilon] \to 0 \tag{3}$$

as $n \to \infty$ and that the probability approaches zero exponentially as $n \to \infty$. Furthermore the bounds on the probability are given explicitly in terms of the network architecture.

This result says that for a large sample it is improbable that there exists a network for which the empirical measure of performance is misleading.

2.4 A self-organizing process

The work on "Stability of a self-organizing process" [4] treats the stability of the Markov chain involved in the self-organizing feature maps of Kohonen. These maps are determined by the effect of a random environment. The values of these maps learn to imitate the environment while also attempting to preserve the neighborhood topology. We give conditions under which two initial states approach each other exponentially fast for all time with

probability one. Thus the initial state does not matter; the environment determines future history.

One way of constructing a Markov chain is as follows. Consider a state space and a sequence $F_1, \ldots, F_n, \ldots$ of independent random functions from the state space to itself. Take an initial point X. The random orbit defined by the iterations $X_n = F_n \cdots F_1 X$ is the Markov chain.

The self-organizing feature map is constructed in this way. Consider a finite subset Λ of the integer lattice $\mathbf{Z}^d$ with ℓ points. The state space is the set of all functions X from Λ to $\mathbf{R}^N$.

Fix a probability measure μ on $\mathbf{R}^N$ and a *shrinking* parameter α with $0 \leq \alpha \leq 1$. Also fix an integer *range* parameter $r \geq 0$. The random functions F_n are defined as follows. For each n choose an independent point ω in $\mathbf{R}^N$ from the probability distribution μ. Then choose the i in Λ that minimizes the distance $|X(i) - \omega|$. Consider the *neighborhood* of i consisting of all j in Λ with $|i - j| \leq r$. Then $F_n X$ is the new state where $F_n X(j) = \alpha\omega + (1-\alpha)X(j)$ for all j in this neighborhood and the other values are unchanged.

One usually takes Λ to be ℓ points constituting a rectangular subset of $\mathbf{Z}^d$. For mathematical investigations it is convenient to take μ to be the uniform measure on a product of N intervals.

The main parameters for the feature map are the integers d and N, representing the dimensions for domain and range, and the range r of the interaction in the domain The number of points ℓ in the domain is another parameter, and finally there is the shrinking parameter α.

Most often $d = N$; however sometimes it is interesting to take more general $d \leq N$.

The case of zero range $r = 0$ is a self-organizing *clustering* process. In this case Λ plays the role of a structureless index set. However in the case $r = 1$ of *nearest-neighbor* interaction the process attempts to preserve topology, and so might be called *topological clustering.*

One picturesque interpartion of the case $d = N = 2$ is when Λ is thought of as a set of cells in the cortex and the region in $\mathbf{R}^2$ is thought of as the retina. The map X from the cortex to the retina develops in such a way that the points in the retina fall in areas of the retina with extensive stimulation, while at the same time nearby cells in the cortex tend to be connected to nearby points in the retina.

We consider the *stability* question, formulated as follows. Take $\alpha > 0$. Take two initial states X and Y and look at their orbits under the same

random sequence of functions. The question is whether when the initial X and Y are sufficiently close there is a non-zero probability that they stay close for all future time.

If X and Y are close, then the i that makes $X(i)$ closest to ω may also be the i that makes $Y(i)$ closest to ω. In that case, $X(i) - Y(i)$ is replaced by $(1 - \alpha)(X(i) - Y(i))$. Thus the two points have been shrunk together. The question is whether this shrinking can persist for all future time.

The ergodic behavior has been studied rigorously in a fundamental paper by Cottrell and Fort. They treat the one-dimensional case $d = 1$ and $N = 1$ with nearest neighbor interaction (range $r = 1$). One interesting feature of this case is that it makes sense to say that a map X is increasing or decreasing. Cottrell and Fort prove that the map is eventually either increasing or decreasing. Once it has reached one or the other status it remains that way. Furthermore for maps of either status they prove convergence to a stationary distribution. (The higher dimensional case has been studied by Ritter and Schulten. Much of their work uses a diffusion approximation.) We prove stability for the one-dimensional case studied by Cottrell and Fort. We also obtain partial results on the higher dimensional case.

2.5 A neural oscillator

One component of the research is modelling of parts of the nervous system in various invertebrates [5]. This has involved the creation of a computer model that can accommodate biologically realistic parameters and a wide variety of neural connections. Members of the experimental group in Neuroscience have been helpful in this enterprise. Brian Smith and Tom Christensen have made suggestions about modelling parts of the insect olfactory system, and Ed Arbas has provided data on the leech heartbeat timing mechanism. One result of the latter contact has been research on neural oscillators. The experimental background is contained in work of Ronald L. Calabrese, James D. Angstadt, Edmund A. Arbas, "A neural oscillator based on reciprocal inhibition," in *Perspectives in Neural Systems and Behavior*, T. J. Carew and D. Kelley (eds.), Alan R. Liss, Inc., N.Y., 1989. The system studied in these experiments was the subject of a preliminary computer analysis. This led to work that gave a better theoretical understanding of the system.

In the simplest version, the oscillator consists of two neurons, numbered 1 and 2. One can make the idealization that the oscillator is described by

four variables, the two membrane voltages v_1 and v_2 and two conductances g_1 and g_2. The system is described by four differential equations for the four variables. The equations for the two voltages v_i and two excitatory conductances g_i are

$$c\frac{dv_i}{dt} = (E_- - v_i)s(v_{i^*}) + (E_+ - v_i)g_i$$

and

$$\tau\frac{dg_i}{dt} = r(v_i) - g_i.$$

Here i ranges from 1 to 2 and labels one neuron, while $i^* = 3 - i$ labels the other neuron. The voltage equations are the current conservation equations for a circuit with two parallel channels. The parameter $c > 0$ is the capacitance. The parameters E_- and E_+ are voltages induced in the two channels by ionic diffusion. One takes $E_- < E_+$ and refers to these as inhibitory and excitatory channels. The conductance $s(v_{i^*})$ in the inhibitory channel is given by a positive increasing function s of the voltage v_{i^*} in the other neuron. (Thus a high voltage in one neuron increases the conductance in the inhibitory channel in the other neuron; the two neurons are mutually inhibitory.) The conductance in the excitatory channel is given by the variable g_i. The equations for the excitatory conductances involve a time constant $\tau > 0$ that governs the rate of their approach to equilibrium. The equilibrium conductance $r(v_i)$ is a positive decreasing function r of the voltage v_i in the same neuron. One wishes to compare the solution of the system of four differential equations with the *discontinuous solutions* of systems of two differential equations. There are two such systems, one defining "slow motion" (in voltage equilibrium) and one defining a "fast motion" (between voltage equilibria). One leaves the slow motion at a "junction point" where the motion cannot continue in voltage equilibrium. Then a fast motion moves the system very rapidly to a "drop point," where the voltage equilibrium is resumed.

3 Network Connectivity

3.1 The trapping transition: communication blockage in a random network

This project, which was carried out by graduate student M. Pokorny in collaboration with Newman and with D. Meiron of Caltech, is primarily percolation theoretic and involves computer simulations of two-dimensional randomly connected bond networks [6]. Let us describe the basic phenomenon under investigation using neural network terminology.

Suppose all active synapses are short range (say nearest neighbor), but only a fraction p of all possible short range synapses are actually active. Suppose further that p is above the percolation threshold p_c, so that long range communication is possible through the infinite cluster of active synapses. Assume that all active synapses in this infinite cluster (but none of those in finite clusters) are suddenly destroyed (e.g. by a massive "seizure" or invasion by a virus which can only be transmitted through the currently active synaptic connections). We ask whether global communication can be restored to the system by activating the currently inactive synapses and utilizing them together with the currently active synapses (which belonged to finite clusters and hence were not destroyed by the seizure/invasion).

It is quite clear that if $p < 1 - p_c$, then the answer to the above question is yes, since then the inactive synapses by themselves percolate. The answer should continue to be yes past $1 - p_c$ until a new critical value k_c is reached; beyond k_c, communication from the origin through undestroyed synapses is blocked or trapped by the destroyed synapses. Let us call the set of sites that can be communicated with from the origin in this situation its "trap." In our simulations we investigated the value of this trapping critical point k_c, and more significantly the issue of whether the trapping phase transition is in the same "universality class" as the conventional percolation transition at p_c. The work was motivated by earlier work on the related issue of "invasion percolation" by Willemsen and Wilkinson and by J. and L. Chayes and Newman.

Our numerical results for two dimensions are as follows. The value of k_c is estimated as about 0.520, which is indeed above $1 - p_c$ (which here is exactly equal to 0.5). The critical exponents were studied by estimating the mean trap size for p above k_c and then using finite size scaling procedures.

The conclusion is that the critical exponents are completely consistent with those occurring in the ordinary percolation phase transition. Thus there does not appear to be a new universality class involved (disagreeing with earlier claims of Willemsen and Wilkinson). L. Kadanoff has noted that this negative result may be of more than passing interest, since it was thought that invasion percolation, being a dynamical model, was capable of generating critical phenomena not obtainable by static percolation models.

3.2 Markov fields on branching planes: the connectivity transition in a layered network

This project, which is part of graduate student C. Wu's thesis research, involves a mixture of percolation and Ising theoretic techniques in the analysis of the connectivity properties of layered networks [7]. The type of network studied is a stack of tree-like graphs (Bethe lattices) which, in addition to their "horizontal" tree graph edges also have "vertical" nearest neighbor edges between the individual layers of the stack. The edges may be thought of as potential synaptic connections.

In the percolation version of the model, only a fraction p of the horizontal synapses and a fraction v of the vertical synapses are active. G. Grimmett and Newman studied the global connectivity properties of the system as these two parameters vary and discovered a pair of transitions: first from local connectivity to large scale connectivity that is concentrated within layers, and then to diffuse large scale connectivity (i.e., in which the original layered nature of the model is not mirrored by the connectivity pathways).

In the continuation of this work by Newman and Wu, it has been shown that this double transition persists for Ising (2-state neuron) and Potts (multi-state neuron) models on such a layered network, as a pair of coupling parameters are varied.

3.3 Ising models and dependent percolation

The relation between ferromagnetic (excitatory) Ising systems and percolation models has been very fruitful, as in the work of [7]. For systems with both inhibitory and excitatory synapses, the relation to percolation is more complicated. In [8], Newman reviews this relation and treats some extra issues that arise for multi-state neurons (Potts models).

3.4 Topics in percolation

This review article article [9] covers recent work in percolation including that of [6, 7, 8].

4 Computer Networks and Computer Performance Analysis

The research supported by this contract has included a substantial component on the performance analysis of stochastically modelled computer networks, and of computer systems in general [10, 11, 12, 13, 14]. Robert Maier has investigated a number of *resource contention models* both theoretically and numerically. Simulations have been performed on a Connection Machine located at the head offices of Thinking Machines in Cambridge. The Connection Machine was made available under a DARPA grant, and accessed *viâ* the national Internet.

The probabilistic models investigated were all models of *complex systems*: systems comprising many loosely cooperating agents, the overall behavior of which is not easily predictable. The prototypical example is a network of computers, which compete for access to a single communication channel. Another is a system in which agents compete for a divisible (rather than discrete) resource, such as computer memory. In both cases competition can get out of hand: the resource may be exhausted, or be poorly apportioned. The mathematical study of such phenomena requires estimates on the first passage times of Markov processes, and large deviation theory is a major tool.

The following subsections review the models investigated, and summarize the papers completed and in preparation.

4.1 Network instabilities

Random access broadcast channels are frequently employed in packet-switched data communication networks. They have two defining characteristics:

- A single logical bus over which most data moves. Data transmissions are usually received by all nodes on the network, and overlapping transmissions can interfere with one another.

- A comparative absence of control signalling, whether out-of-band or in-band. Nodes transmit randomly and more or less independently, according to some protocol.

Since the nodes are independent, in the event of packet collisions they must implement their own retransmission strategies. They must buffer blocked or unsuccessfully transmitted packets and attempt to retransmit them later.

Many protocols for controlling random access broadcast channels are *unstable*: under conditions of heavy load the packet collision rate may suddenly rise to an unacceptable level. But determining theoretically whether any given random-access protocol is unstable is a vexing problem. Such protocols as CSMA/CD [Carrier Sense Multiple Access/Collision Detect] have been repeatedly studied, both analytically and numerically, but the results of the studies have been of limited practical applicability. This is in part because much work has concentrated on *infinite-user* models: network models in which the total rate of packet arrivals is finite, but the number of network nodes is infinite.

Several authors, beginning with Kleinrock and Lam, have explored the way in which such instabilities appear in stochastically modelled networks as N, the number of nodes, tends to infinity. Studies of large-N networks have revealed why heavily loaded networks become unstable: the associated Markov chain exhibits two points of stable equilibrium. One corresponds to a desirable high-throughput, low-contention state, and the other to an undesirable low-throughput, high-contention state. The sudden degradation in network performance corresponds to a transition from the former to the latter.

Robert Maier's research [10] evaluated the performance of computer networks in the large-N limit, but a large-N limit different from that used in the conventional infinite-user model. As $N \to \infty$, the packet retransmission rates on each individual node, as well as the packet arrival rates, were taken proportional to N^{-1}. This choice of scaling allowed the use of Ventcel-Freidlin theory, the theory of the large deviations (*i.e.*, leading-order fluctuations) of Markov chains. If the time until performance degradation is denoted τ, then according to the theory its expectation has leading-order asymptotics

$$\mathbf{E}\{\tau\} \sim C N^{\alpha} e^{N S_0}, \quad N \to \infty$$

in which S_0 can in principle be computed exactly.

The Ventcel-Freidlin formalism was applied to two random access systems: a time-slotted ALOHANET model, and an unslotted persistent CSMA/CD model. In both cases S_0 was computed. The retransmission protocol used in the latter model closely resembled that used by Ethernet, so the computation of S_0 is relevant to the stability of real-world computer networks.

4.2 Local contention models

The network models reviewed in the last subsection exhibited a comparatively simple contention for resources. There was only a single resource: the network bus. All agents competed for it, and competition was global.

Robert Maier is now investigating, both theoretically and by means of simulations, a more realistic model of *local* contention [11]. Consider a Euclidean or hypercubic grid of processors, each equipped with its own local memory. Suppose that each processor is running its own program, and that there is minimal interprocessor communication. Any given processor may, however, from time to time require more memory than it has available. It may attempt to access, and use, the memory resources of its nearest neighbors.

This sort of contention for resources is purely local, and the resources are as distributed as the agents. But the simplest scheme for resolving collisions is the same as in the global case: 'colliding' nodes (those that wish to use the same memory at the same time) resolve the deadlock by backing off a random amount of time and trying again. Such a random-access protocol can be quite efficient. But under conditions of heavy load, *i.e.*, frequent requests for adjacent memory, it may give rise to bistability. As in the global case, collisions and retries may escalate suddenly to an unacceptable level.

The methods of Ventcel and Freidlin are not appropriate to this problem; although it can be thought of as a Markov chain problem, the dimensionality of the state space is very large. In fact it must be approached largely through simulations. Simulations are being conducted on the Connection Machine at Thinking Machines' offices in Cambridge. The Connection Machine is being accessed *viâ* the NSFNET, and Sun and other workstations at the University of Arizona are being used as front ends.

4.3 Memory exhaustion

Another problem investigated by Robert Maier was that of memory exhaustion in a system with only two competing agents [12]. The problem considered was that of two stacks evolving in a bounded region. This is one of the simplest problems in the area of dynamic data structures, but a full analysis requires surprisingly deep mathematics. It was originally proposed by Knuth, and was investigated further by Flajolet and Louchard. Ventcel-Freidlin theory, it turns out, allows the treatment of the case in which the mobility of the stacks is allowed to vary with height.

The basic questions are as follows. Suppose one allocates a contiguous block of m cells of memory, and allows two stacks to evolve (*i.e.*, randomly grow and shrink) within it. The two stacks begin on opposite sides of the block; one grows upward and the other downward. How long will it be before they collide? And at the time of collision, what will their sizes be? It is *asymptotic* estimates, valid as $m \to \infty$, that are desired.

The time to collision and the final stack sizes are of course random variables. Their distributions will depend on the probabilities assigned to the possible evolutionary histories of the stack system. Knuth suggested that at each time step, there should be probability p of each stack increasing by one in size, and probability $1/2 - p$ of each stack's height decreasing. This choice defines a Markov chain on the space of states of the two-stack system, which is parametrized by the two stack heights.

The initial state of the system is $(0, 0)$, and the set of final states F is

$$\{(j, k) \in Q_m \mid j + k = m\}.$$

In terms of the Markov chain, the basic questions are

- How many operations take place before a state in F is reached?
- Which state will it be?

The answers are probabilistic, and depend markedly on p.

The $p < 1/4$ case is the most realistic, and the hardest to analyse. With this choice of p the stacks will be biased toward contraction, and the mean time to collision will grow exponentially rather than polynomially in m. Flajolet made an interesting discovery: in this case, the limiting distribution

on F is *uniform*. That is to say, if m is large and the stacks are biased toward contraction, when the stacks finally collide they are as likely to be large as small. This has obvious implications for the design of algorithms. In particular it would seem to indicate that the *shared* storage method considered here, with the stacks sharing a block of size m, is usually far more efficient than a *separate* arrangement, with the stacks confined to their own regions of size $m/2$. The latter scheme would run out of memory much earlier.

Robert Maier showed that the uniform distribution over F is an *artifact*, attributable to Knuth's choice of random process. Traditionally the behavior of the two stacks is assumed to be *independent of their size*: p, the probability of an insertion into either stack, is independent of state. It is reasonable to allow more general behavior: the probability of insertions into and deletions from each stack could depend on the height of the stack as a fraction of m. In particular one can let the insertion probability be $(1 - g(x))/4$, and the deletion probability be $(1 + g(x))/4$, in which x is the height of the stack, as a fraction of available memory, and $g(\cdot) \geq 0$ is some sufficiently smooth function defined on $[0, 1]$. $g(x)$ measures the extent to which deletions from a stack predominate over insertions.

If g is not constant, the limiting distribution on F will not be uniform. Two possibilities deserve mention: either the limiting distribution will be localised at $(m/2, m/2)$, or it will be concentrated at $(0, m)$ and $(m, 0)$. The former occurs if $g' > 0$, in which case the stacks become more biased toward contracting as they grow. It also occurs if $g(0) = 0$, in which case the bias toward contraction disappears at low stack height. The latter typically obtains if $g' < 0$.

Besides the limiting distribution on F, the distribution of τ, the number of operations that take place before F is reached, was studied. In the case of constant $p < 1/4$, which corresponds to constant $g(\cdot)$, Flajolet proved combinatorially that its distribution is asymptotic to that of an *exponential* random variable, with mean roughly $O(((\frac{1}{2} - p)/p)^m)$. It was shown that this phenomenon occurs very generally: for *any* sufficiently smooth $g(\cdot) > 0$, there are constants C_0, α and S_0 such that $\tau / C_0 m^\alpha e^{m S_0}$ is asymptotically exponential with unit mean. Explicit formulæ for C_0, α and S_0 in terms of $g(\cdot)$ were derived.

4.4 Weighted evolutions of data structures

Robert Maier's research included some other work on stochastically modelled data structures. The emphasis was on the asymptotics of the cost of a sequence of operations on a data structure, rather than on the appearance of instabilities or other undesirable behavior. The same Ventcel-Freidlin theory proved useful, however.

A number of authors had previously derived asymptotic expressions for the average cost of n operations on such data structures as priority queues and dictionaries. The expressions depended on (1) the implementation of the data structure, and (2) the probability measure over random sequences of operations (insertion, deletion, and queries of various sorts) used in computing the average.

In the present investigation [13] an *equiprobability of histories* was assumed: all possible sequences of alterations of the data structure were taken as equiprobable. This included alterations consequent on the insertion of a datum, on the deletion of a datum, and on accessing the structure to query it or to alter a datum in some way without removing it. The cases of list and d-heap implementations of priority queues were treated, and the case of list implementations of linear lists and dictionaries. In the case of dictionaries, an arbitrary number of query types were allowed.

In this framework, results on list implementations had been obtained previously by the combinatoric techniques of Flajolet and the more probabilistic method of Louchard. The present treatment extended theirs by covering heap implementations as well as lists. Much more importantly, it brought to bear the powerful formalism of path integration. This formalism originated in physics, and can be viewed as a user-friendly form of Ventcel-Freidlin theory.

The path integral formalism makes possible a very general analysis of equiprobable structure histories: so long as the structure implementations are of the simple list or heap form, the allowed operations can be considerably more sophisticated than mere insertions or deletions. The assumption of strict equiprobability of histories is also unnecessary: one can easily treat models in which histories are differently weighted, with certain operations taken as more likely than others. Such models are particularly difficult to handle combinatorially.

The applicability of the path integral method is not so much restricted by the choice of datatype as by the choice of implementation. In general

this method, with its assumption of a comparatively small state space, can be used to treat implementations that "uniquely represent their data." Such *implicit implementations* allow at most one representation, up to order isomorphism, for any quantity of internally stored data.

It was shown that for such implementations the *asymptotic determinism* found by Louchard occurs very generally: in the limit of long histories, the most likely evolutions of the data structure are those that cluster tightly around a deterministic path. In consequence the integrated space and time costs, when normalized, converge as $n \to \infty$ to deterministic values. In the case of list implementations the limiting costs are quadratic in n, so that expected costs increase in the limit as fast as worst-case costs. But for heaps the expected costs turn out to increase rather less rapidly: the integrated space cost as $n^2/\sqrt{\log n}$, and the integrated time as $n \log n$. In expectation the spatial cost differs markedly from its worst-case value, which is quadratic in n.

This unusual phenomenon — the clustering of integrated costs around deterministic values — probably does not occur in real-world databases. Its failure to appear is a sign that in the real world, data structure histories are far from being equiprobable.

4.5 Stochastic orderings and mean extremes

An additional topic investigated by Robert Maier is more theoretical, but has applications to queueing theory and computer performance analysis. In collaboration with Peter Downey of the University of Arizona, he has investigated the relationship between stochastic orderings of random variables and the growth rate of their mean extremes[14].

Consider n independent copies of a non-negative random variable X. Denote by $X_{(n)}$ the maximum of these copies. It is a random variable itself, and is called the *maximum order statistic*, or the *extreme* of the n-element sample. If X is unbounded above, then as $n \to \infty$ the mean extreme $\mathbf{E}\{X_{(n)}\}$ will also tend to infinity. Its growth rate is of interest.

If X is an exponential random variable, it is easy to check that $\mathbf{E}\{X_{(n)}\}$ grows logarithmically in n. Maier and Downey showed that it grows no faster than logarithmically if and only if X is suitably bounded by an exponential: in particular, if and only if it is stochastically dominated by an exponential variate.

This result is really only a special case of a new result on stochastic orderings. One of these orderings is *variability order*. Variability ordering is a partial order on non-negative random variables; one says that $X \leq_c Y$, or that X is *less variable* than Y, if $\mathbf{E}\{h(X)\} \leq \mathbf{E}\{h(Y)\}$ for all increasing convex functions $h: \mathbf{R}_+ \rightarrow \mathbf{R}_+$. Another ordering on non-negative random variables is the *mean extreme ordering*, defined as follows: $X \leq_o Y$ means that the mean extremes of X are all less than or equal to the corresponding mean extremes of Y.

Maier and Downey showed that the orderings $\leq_c$ and $\leq_o$ are very closely related: $X \leq_c Y$ implies $X \leq_o Y$, and $X \leq_o Y$ implies $X \leq_c CY$ for some universal constant C. So X is bounded in variability ordering by Y if and only if its mean extremes grow no more rapidly, as $n \rightarrow \infty$, than those of Y. But it is easy to show that X is bounded in variability ordering by an *exponential* variate Y if and only if X is stochastically dominated by Y. So the above result follows.

4.6 Phase-type distributions

This work [15] contains an algorithm that constructs, from a given rational function, a Markov chain whose absorption-time distribution has the rational function as generating function. This provides an algebraic proof of C. O'Cinneide's recent characterization of discrete phase-type distributions. The algorithm is based on an automata-theoretic algorithm due to M. Soittola.

Moreover the characterization of continuous phase-type distributions follows from the discrete characterization. In conjunction with the discrete-time algorithm, this engenders an algorithm for constructing a Markov process representation for any distribution of continuous phase-type. This work succeeded in tying together the theory of Markov chains with absorption and the theory of finite automata.

5 Bibliography

References

[1] William G. Faris and Robert S. Maier, "Associated random variables and reliability of neural networks," *J. Neural Network Computing*, to appear.

[2] Robert M. Burton and William G. Faris, "Reliable evaluation of neural networks," *Neural Networks*, to appear

[3] Robert M. Burton and William G. Faris, "Large deviations for neural network loss functions," preprint in preparation.

[4] Robert M. Burton and William G. Faris, "Stability of a self-organizing process," *Abstracts American Mathematical Society* **11** (1990), 462, preprint in preparation.

[5] William G. Faris, "Singular perturbation analysis of a neural oscillator," preprint in preparation.

[6] Martin Pokorny, Charles M. Newman, and Daniel Meiron, "The trapping transition in dynamic (invasion) and static percolation," *J. Physics A: Math. Gen.* **23** (1990), 1431–1438.

[7] Charles M. Newman and Chuntao C. Wu, "Markov fields on branching planes," *Probability Theory and Related Fields*, to appear.

[8] Charles M. Newman, "Ising models and dependent percolation," in *Topics in Statistical Dependence* (H. W. Block, A. R. Sampson, and T. H. Savits, eds.), IMS Lecture Notes-Monograph Series, to appear.

[9] Charles M. Newman, "Topics in percolation," in *Proceedings of the 1989 Summer Seminar on Mathematics of Random Media* (W. Kohler and B. White, eds.), Lectures in Applied Mathematics, American Mathematical Society, to appear.

[10] Robert S. Maier, "The first exit times of multiaccess broadcast channels," preprint.

[11] Robert S. Maier, "Local resource contention: A performance analysis," preprint in preparation.

[12] Robert S. Maier, "Colliding stacks: A large deviations approach," preprint.

[13] Robert S. Maier, "A path integral approach to data structure evolution," preprint.

[14] Peter J. Downey and Robert S. Maier, "Stochastic orderings and extreme moment growth," preprint.

[15] Robert S. Maier, "Algebraic construction of phase-type distributions," preprint.